SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra

Bailey, CM; Sweet, Steve; Cunningham, Debbie; Zeller, M; Heath, John; Cooper, Helen

DOI:
10.1021/pr800917p

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Bailey, CM, Sweet, S, Cunningham, D, Zeller, M, Heath, J & Cooper, H 2009, 'SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra', *Journal of Proteome Research*, vol. 8, no. 4, pp. 1965-1971. https://doi.org/10.1021/pr800917p

research articles Journal of proteome research

SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra

Christopher M. Bailey,[†,‡] Steve M. M. Sweet,[†,‡] Debbie L. Cunningham,[†,‡] Martin Zeller,[§] John K. Heath,[†,‡] and Helen J. Cooper*[,‡]

School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, and Thermo Fisher Scientific, Hanna-Kunath-Str.11, 28199, Bremen, Germany

Received October 29, 2008

Recently, software has become available to automate localization of phosphorylation sites from CID data and to assign associated confidence scores. We present an algorithm, SLoMo (Site Localization of Modifications), which extends this capability to ETD/ECD mass spectra. Furthermore, SLoMo caters for both high and low resolution data and allows for site-localization of any UniMod post-translational modification. SLoMo accepts input data from a variety of formats (e.g., Sequest, OMSSA). We validate SLoMo with high and low resolution ETD, ECD, and CID data.

Keywords: phosphorylation • phosphopeptide • phosphoproteomics • site localization • Ascore • mass spectrometry • post-translational modifications • bioinformatics

Introduction

Large-scale mass spectrometric identification of phosphopeptides and phosphoproteins, termed phosphoproteomics, has now become routine.[1−5] Until recently these analyses relied heavily on manual validation to confirm correct site localization.[6] This approach is both time-consuming and labor intensive, making it impractical for large data sets. To reduce the requirement for manual analysis, algorithms have been developed to automate site localization.[6−9] These approaches use statistical models for assessing site localization, and can be used for computational analysis of large data sets in a short period of time. The algorithm takes the peptide sequence identified from a database search, compiles a list of all possible phosphorylation sites (or combinations of sites), and from this generates a list of predicted ions. These ions are then matched against the mass spectra from which the peptide was initially identified in order to identify the number of matched ions, and in particular the number of site-determining ions. Site-determining ions are those ions which are unique for a particular modification site/combination of modification sites.

While these algorithms represent a major step forward in the development of tools to analyze large scale experimental data sets generated from LC−MS/MS, to date their usage is limited to peptides fragmented by collision-induced dissociation (CID). Recently, the radical-driven fragmentation techniques electron transfer dissociation (ETD)[10] and electron capture dissociation (ECD)[11] have been applied to the large-scale analysis of phosphorylation.[4,5,12] These techniques are particularly well-suited to the analysis of phosphopeptides, as the labile phosphate moiety is retained on peptide backbone fragments, in contrast to fragmentation by CID in which loss of the modification is the dominant pathway.[13] An additional limitation of the existing algorithms is that they accept peptide identification output only from specific search engines, for example, Sequest for Ascore and Mascot for MSQuant-incorporated localization.[6,7]

To address these limitations, we sought to generate a new site localization tool, based on the Ascore algorithm, but capable of addressing a wider range of problems. In particular our aims were to generate a tool which: (i) allows analysis of data obtained using both CID and ETD/ECD fragmentation methods; (ii) caters for both high and low resolution fragmentation data; (iii) enables data to be read into the tool from a variety of formats, using an extensible scheme; (iv) allows analysis to be performed for a variety of modification types.

Our algorithm, SLoMo (Site Localization of Modifications), allows calculation of hypothetical ions based on either CID or ETD/ECD. SLoMo accepts the generic pepXML input format[14] and incorporates options for searching for any modification found in the UniMod database.[6,15] Finally, we incorporate a common database back-end allowing easy access to the source data for advanced users, and increased speed for searches on preproduced databases. We validate SLoMo by comparing it to Ascore (for CID data, where both algorithms will produce results), through the use of synthetic phosphopeptide libraries and by manual validation of SLoMo localizations from *in vivo* phosphorylated proteins. We demonstrate that SLoMo can successfully localize phosphorylation sites from low resolution ETD and CID data, and from high resolution ETD and ECD data. The application of SLoMo to other modifications is illustrated by localization of sites of methionine oxidation. Finally, SLoMo analysis of OMSSA output (phosphorylation) and Sequest output (oxidation) is shown.

* To whom correspondence should be addressed. Helen J. Cooper, School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. Telephone: +44 (0)121 414 7527. Fax: +44 (0)121 414 5925. Email: H.J.Cooper@bham.ac.uk.

† CRUK Growth Factor Group.

‡ University of Birmingham.

§ Thermo Fisher Scientific.

10.1021/pr800917p CCC: $40.75

Published on Web 03/10/2009

Experimental Section

Spectrum Preprocessing. Given an MS/MS spectrum generated for a peptide of precursor mass m, and a charge z, the mass spectrum is preprocessed to remove peaks corresponding to intact precursor ions and charge-reduced products (in ETD/ECD mass spectra) and selected neutral losses (in CID mass spectra) prior to binning the spectrum into 100 m/z windows. SLoMo incorporates a tolerance t for region removal and utilizes the general ppm error value e (also used in the generated ion-to-mass spectrum matching part of the algorithm) such that for ECD/ETD data peaks within the region R are removed for:

$$\frac{m + Hi - e}{i} \leq R \leq \frac{m + Hi + \frac{t}{i} + e}{i}$$

where i is the charge state of the reduced precursor ion, ($1 \leq i \leq z$), and H is the mass of a proton.

For CID data, peaks within region R are removed for:

$$\frac{m + Hz - e - n}{z} \leq R \leq \frac{m + Hz + e + \frac{t}{z} - n}{z}$$

where n is the mass of the neutral loss entity, for example, $n =$ 97.977 Da (H_3PO_4) for phosphopeptides. Potential neutral losses are obtained by looking up data on the modification being analyzed in UniMod. Each neutral loss is considered alone or with additional neutral loss of water, and the region corresponding to the neutral loss of water alone is also removed.

Ion List Generation/Filtering. Where ions are to be matched against mass spectra generated from ECD/ETD events, c and z ions N-terminal to a proline residue are not generated. (N-terminal proline c and z ions are rarely observed in ECD/ETD mass spectra[16]). Where there are multiple modifications of different types (e.g., phosphorylation and acetylation) on a single peptide only one type of modification will be used to generate combinations of modification sites: all other modifications are considered to be correctly localized by the search engine used to generate the peptide identifications, and remain fixed.

Modification Matching. Modifications are mapped to UniMod records by matching modification information stored within the pepXML file to the UniMod database. Matches are made based on the amino acid identified as modified by the initial protein database search and the mass difference of the assigned modification. Specifically, the modification mass must be within 10^{-4} Daltons of the modification mass recorded in UniMod, and the amino acid must be specified in UniMod as a site capable of having the modification present. For output from the OMSSA database search algorithm,[17] where conversion to pepXML is currently unavailable, a user editable list is provided which links the name given to the modification by OMSSA (e.g., phosphorylation) to the UniMod ID number for that modification (in the case of phosphorylation: 21).

Statistical Calculation. Ascore uses a cumulative binomial probability model to calculate peptide and final Ascores:

$$P = \sum_{k=n}^{N} \binom{N}{k} p^k (1 - p)^{N-k}$$

where N is the number of trials (in this case the number of potential ions generated by the peptide under examination), n is the number of successes (the number of times an ion was matched to a peak in the mass spectrum), and p is the probability of a random match between an ion and a peak. Ascore uses a tolerance of ± 0.5 m/z for matching a peak to an ion. This means that for a given peak depth of n peaks per 100 m/z window the probability of a match by chance is:

$$p = \frac{n}{100}$$

However, in SLoMo, the tolerance is expressed in parts per million rather than in m/z units. Thus the probability becomes:

$$p = \frac{x_{\min} + x_{\max}}{2} \times \frac{2e}{1000000} \times \frac{n}{100}$$

where $x_{\min}$ is the value (m/z) of the lowest m/z peak in the mass spectrum, $x_{\max}$ is the value (m/z) of the highest m/z peak in the mass spectrum, and e is the tolerance in parts per million.

In order to overcome computational issues associated with calculating the binomial probability for large number of trial events (which can lead to overflow and/or underflow issues during calculation), SLoMo utilizes a cumulative Poisson distribution when the number of trial events is such that tests for underflow and overflow issues in the binomial calculation are true. When this is the case, the following equation is used to calculate the probability score:

$$P = \sum_{k=n}^{N} \frac{e^{\lambda} \lambda^k}{k!}$$

where $\lambda = np$.

Test Datasets. 1. Ascore Data Set. A test data set of 135 phosphoserine containing peptides was downloaded from http://ascore.med.harvard.edu/examples/sp.zip to test against SLoMo and Ascore (run online at http://ascore.med.harvard.edu/ascore.php.

2. Synthetic Phosphopeptide Libraries. Synthetic phosphopeptide libraries were synthesized by Alta Biosciences (Birmingham, UK). The libraries were based on the peptide sequences: GPSGxVpSxAQLx[K/R] and SxPFKxpSPLxFG[K/R], where x is from ADEFGLSTVY. Each library therefore contains a mixture of 2000 phosphopeptides. Each phosphopeptide library was either fractionated by SCX chromatography or analyzed directly by LC−MS/MS.

3. Whole-Cell Lysates. Mouse fibroblast NIH 3T3 cells were cultured at 37 °C, 5% CO_2 in Dulbeccos Modified Eagle Medium (Invitrogen) supplemented with 2 mM L-Glutamine (Invitrogen), 0.1 mg/mL streptomycin, 0.2 U/mL penicillin (Sigma) and 10% v/v donor bovine serum (Invitrogen). Following serum starvation in media containing 0.1% serum for 18 h, cells were treated with 2 mM sodium pervanadate for 20 min, prior to lysis. Cells were lysed by sonication in ice-cold urea lysis buffer (17 mM HEPES pH 8, 7.65 M urea, 1 mM Na_3VO_4, 50 mM NaF, 25 mM β-glycerophosphate and 1 tablet of complete mini protease inhibitor cocktail (Roche Diagnostics) for every 10 mL of buffer).The lysates were reduced (8 mM DTT) and alkylated (20 mM iodoacetamide) in 50 mM ammonium bicarbonate. The lysates were diluted to 4 M urea, acetonitrile (10% by volume) and endoproteinase Lys-C were added (Sigma; 1:400 enzyme:protein) and digestion was allowed to proceed at 37 °C for 5 h. The lysates were then further diluted to 1 M urea, trypsin (Trypsin Gold; Promega, Madison, WI) was added (1:100 enzyme:protein) prior to overnight digestion at 37 °C. Peptides were desalted and phosphopeptides were enriched using TiO_2 affinity.[18] Bulk enrichment was carried out in an

Eppendorf-type tube, rather than small columns, however the rest of the protocol was as previously described.[19] Both the phosphopeptide-enriched eluate and the flow-through (unbound fraction) were retained for MS analysis.

Liquid Chromatography Tandem Mass Spectrometry (LC−MS/MS). CID and ECD Data. Online liquid chromatography was performed by use of a Micro AS autosampler and Surveyor MS pump (Thermo Electron, San Jose, CA). Peptides were loaded onto a 75 μm (internal diameter) Integrafrit (New Objective, USA) C8 resolving column (length 10 cm) and separated over a 40 min gradient from 0% to 40% acetonitrile (Baker, Holland). Peptides eluted directly (~350 nL/min) via a Triversa nanospray source (Advion Biosciences, NY) into a 7 T Thermo Finnigan LTQ FT mass spectrometer (Thermo Electron), where they were subjected either to data-dependent CID or ECD. CID and ECD parameters were approximately as previously described.[19,20]

ETD Data. Online liquid chromatography was performed by use of a Micro AS autosampler and Surveyor MS pump (Thermo Fisher Scientific, San Jose, CA). Peptides were loaded onto a C18 trapping column (100 μm internal diameter, 2 cm length, nanoseparations, The Netherlands) and separated on a C18 analytical column (75 μm inner diameter, length 10 cm, nanoseparations, The Netherlands) over a 30 min gradient from 0% to 35% acetonitrile (Fisher Scientific, USA). Peptides eluted directly (~300 nL/min) into a LTQ Orbitrap XL ETD mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Electron transfer dissociation was induced in the LTQ and ETD fragment ions were either detected in the LTQ or in the Orbitrap mass analyzer.

ETD reaction time was set to 100 msec for all charge states. Singly charged ions were rejected. Target value settings were 5×10^5 for FT full scans, 1×10^5 for ETD with Orbitrap detection and 1×10^4 for ETD with LTQ detection.

Data Analysis: OMSSA Database Search Algorithm. DTA files were created from the raw data using Bioworks 3.3.1 (Thermo Fisher Scientific Inc.). For the whole-cell lysate data, the OMSSA browser was employed to search the DTA files against a concatenated database consisting of the mouse IPI database (Version 3.40) and the reversed-sequence version of the same database. The synthetic phosphopeptide libraries data were searched against a concatenated database consisting of 32 400 NCBI *Drosophila melanogaster* sequences and the library sequences with either K or R added to the N-terminus (to create reversed versions of the same mass).

OMSSA settings for the whole-cell lysate high mass accuracy ETD and ECD searches were as follows. Enzyme: trypsin. Peptide m/z tolerance: 1.1 ±. MS/MS m/z tolerance: 0.02 ±. Mis-cleavage allowed: 2. Fixed modifications: carbamidomethyl C. Variable modifications: acetylation of protein N-terminus, oxidation of M, phosphorylation of S, T, Y. Product ion types to search: c, z, y. E-value cutoff: 50. Allow N-terminal Met cleavage: yes. Allow elimination of charge-reduced precursors in spectrum: yes. Precursor charge-state detection: read from input file data. Alterations to the above settings for the various searches are detailed below. OMSSA settings for the synthetic peptide ECD search: Mis-cleavage allowed: 1. Variable modifications: phosphorylation of S, T, Y. OMSSA settings for the low mass accuracy synthetic peptide CID search: Peptide m/z tolerance: 0.02 ±. MS/MS m/z tolerance: 0.8 ±. Product ion types to search: b,y. OMSSA settings for the low resolution ETD search: Peptide m/z tolerance: 0.02 ±. MS/MS m/z tolerance: 0.8 ±.

OMSSA results were filtered to allow only the top scoring identification (sequence and site of modification) per DTA. The results were then filtered by precursor mass error (in ppm) and e-value to obtain a false-discovery rate for phosphopeptides lower than 2% for each search (FDR = reverse hits/forward hits × 100).

SLoMo settings for the various data sets are detailed below. All searches: tolerance = 4 (window for removing precursor and neutral-loss peaks); doubly charged fragment ions were allowed for triply charged and higher charge-state precursors. ECD and ETD data: c, z, z-prime and y ions; fragment ion tolerance of 13 ppm (high resolution) or 400 ppm (low resolution ETD data). CID data: b, y ions; fragment ion tolerance of 400 ppm.

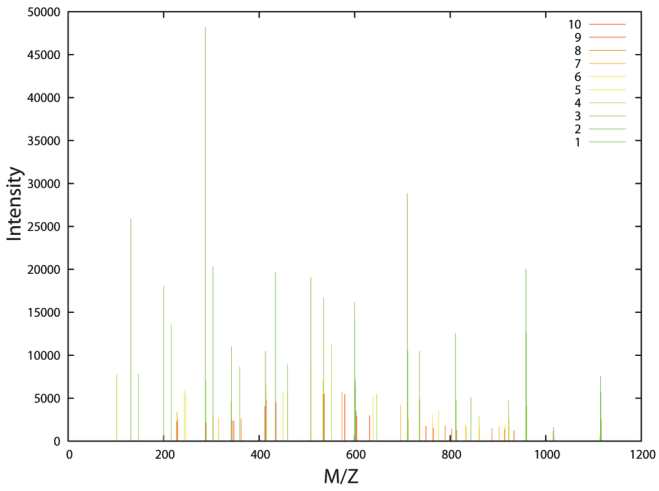
Data Analysis: Sequest Database Search Algorithm. Bioworks (3.3.1) was employed both to generate DTA files from ECD analysis of mouse whole-cell lysate (flow-through from the TiO_2 enrichment described above) and to carry out a database search using the Sequest algorithm.[21] The database searched was the mouse IPI database, as described above. Search parameters were similar to those used above, with the exception that protein N-terminal acetylation was not available as a variable modification. The search results were filtered as above, with the Sequest Xcorr score in place of the OMSSA e-value. The Trans-Proteomic Pipeline (Seattle Proteome Center) was employed to generate a pepXML file. This file was then used by SLoMo to localize methionine oxidation. SLoMo search parameters were as described for ECD above.

Results

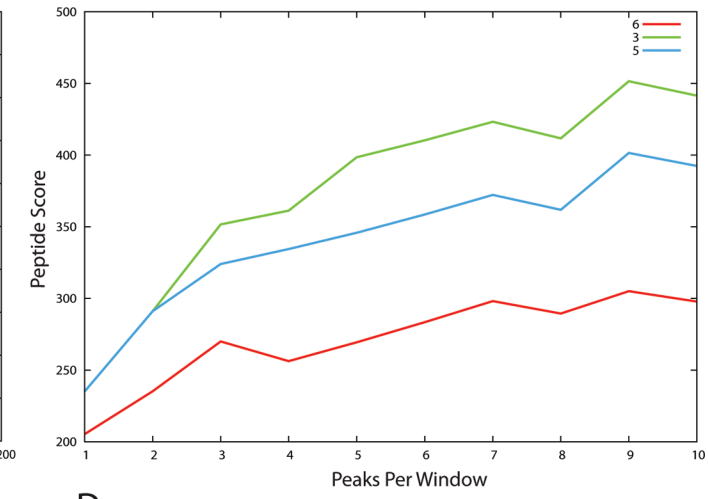
Algorithm. SLoMo is modeled on the Ascore algorithm, which is discussed in detail by Beausoleil et al.[6] We therefore present here an overview of the algorithm with particular attention paid to the areas that have been enhanced.

Assignment of a site localization and its associated score begins with an analysis of the MS/MS spectrum for the peptide in question. The mass spectrum is first cleared of all peaks corresponding to precursor ions, charge-reduced ions, and neutral loss ions according to the dissociation method being used, that is, precursor ions and charge-reduced ions for ETD/ECD and neutral loss ions for CID. The mass spectrum is split into regions, or windows, of 100 m/z width. From each of these regions the most intense i peaks are selected, where i is an integer between 1 and 10. For example, a list generated at 5 peaks per window contains the 5 most intense peaks in each 100 m/z window. Once lists of peaks have been determined for each peak depth, the software then generates a list of all potential modification sites by examining the peptide sequence determined from the mass spectrum via a protein database search algorithm (e.g., Sequest, OMSSA etc.). At this point our new algorithm deviates from the original. Through use of the information available in pepXML data files, it is possible to see all combinations of residues and mass differences considered by the protein database search algorithm when searching for modified peptides. These data are read by SLoMo and used to determine the potential modifications for which it can search. For OMSSA output, SLoMo parses the file to detect all modifications found. For each modification, the corresponding UniMod record is found, and the user prompted to select which amino acid(s) should be considered as possible sites for this modification. Once this step is complete, SLoMo allows the user to select from any of the modification types detected to perform site localization. If there is more than one type of modification

A



B



C

Top hit: IDIs*PSTFRK

Ion	m/z	Diff
c1	131.118	0.000
c3	359.229	0.000
c6	710.313	0.000
c7	811.360	0.001
c8	958.429	0.000
c9	1114.530	0.001
y8_2+	508.253	0.000
y8	1015.498	0.003
y6	735.415	0.001
y5	638.363	0.001
y4	551.331	0.000
y3	450.283	0.000
y2	303.214	0.000
y1	147.113	0.000
z4	535.312	0.000
z3	434.264	0.000
z2	287.196	0.000
z1	131.095	0.000

Second hit IDIsPS*TFRK

Ion	m/z	Diff
c1	131.118	0.000
c3	359.229	0.000
c6	710.313	0.000
c7	811.360	0.001
c8	958.429	0.000
c9	1114.530	0.001
y8_2+	508.253	0.000
y8	1015.498	0.003
y4	551.331	0.000
y3	450.283	0.000
y2	303.214	0.000
y1	147.113	0.000
z4	535.312	0.000
z3	434.264	0.000
z2	287.196	0.000
z1	131.095	0.000

D

```
Scan 1028, IDIsPSTFRK,
1 mods, (415.210170666667)
Trimming precursor masses from spectrum
Regions Removed:
       1243.600 - 1247.630,
        622.304 - 624.319,
        415.205 - 416.548
Best Peaks calculated for spectrum
SS_112607_ECD_SCXA10_14.1028.1028.3
Top Hit 3
Second Hit 5
Optimum difference at 5 peaks per window
Predicted top peptide IDIs*PSTFRK
Predicted second peptide IDIsPS*TFRK
Final Score: 41.9587656505306
```

Site Determining Ions:
c4, c5, $c4^{2+}$, $c5^{2+}$, y6, y5, $y6^{2+}$, $y5^{2+}$,
z6, z5 , $z6^{2+}$, $z5^{2+}$, z′6, z′5, $z'6^{2+}$, $z'5^{2+}$

Figure 1. Example of SLoMo output: SLoMo output for the phosphopeptide IDISPSYFRK. (A) Preprocessed MS/MS spectrum after precursor removal and generation of peak-per-window list. Color represents the ranked intensity of peaks within each window (1 = most intense). (B) Modification scoring output. Each line shows the score for a different site of modification. In this example, the scores for putative phosphopeptide IDISPSY*FRK at each peak depth are shown in red. The scores for putative phosphopeptide IDISPS*YFRK are shown in blue and for putative phosphopeptide IDIS*PSYFRK are shown in green. (Note that the algorithm labels the N-terminal amino acid as "0", hence phosphopeptide IDIS*PSYFRK is labeled "3" and so on). (C) Comparison of site-determining ions. In this example, the highest scoring peptides were IDIS*PSYFRK and ISISPS*YFRK. The two lists show all the ions observed within the MS/MS spectrum at the optimum peak depth, i.e., at which the largest difference in scores is first observed, (defined as 5 in this case). Ions highlighted in red are site-determining. "Diff" indicates mass error (in *m/z*). (D) (Top) Program output from SLoMo describing its actions at each stage of the analysis process in addition to the final score for the localization. (Bottom) All possible site-determining ions for the two highest scoring possibilities in this example.

present in a peptide, the modifications not selected for site localization are assumed to have been correctly localized by the protein database search algorithm.

Once a modification type has been selected, all potential combinations of modification sites within a peptide are calculated, and for each combination a list of hypothetical ions is generated. SLoMo allows different ion sets to be generated. In common with Ascore, lists of b and y ions are generated for CID data. However, SLoMo can also generate any other combination of ion sets, so that mass spectra containing c, y and z• ions generated by ECD/ETD, may be searched. The SLoMo ion generator can produce ion lists for most common ion types, and those ion types can be combined together in any way. Likewise searching for multiply charged ions is also entirely optional and can be turned on or off. An established feature of ECD and ETD is hydrogen transfer between complementary fragments, resulting in c• and z′ ions.[22] The resulting ź ions are commonly observed and can therefore optionally be used in site localization by SLoMo.

The ion lists generated by SLoMo are matched against the peak lists (i.e., 1 to 10 peaks per window). In contrast to the Ascore algorithm, the maximum permitted difference between

Table 1. SLoMo Validation Using Synthetic Phosphopeptide Libraries and Phosphopeptides Enriched from Whole-Cell Lysates[a]

		number of localized sites (errors and multiple top hits)			
data type	number of mass spectra	score: ≥19	15−19	10−15	<10
Synthetic peptides (x is from ADEFGLSTVY)					
CID GPSGxVpSxAQLx[K/R]	487	353 (0)	67 (0)	40 (2)	27 (23)
CID SxPFKxpSPLxFG[K/R]	420	350 (0)	26 (1)	26 (1)	18 (11)
ECD GPSGxVpSxAQLx[K/R]	202	180 (0)	2 (0)	0 (0)	20 (20)
ECD SxPFKxpSPLxFG[K/R]	406	356 (4)	4 (0)	0 (0)	46 (46)
Whole-cell lysate (manually checked localizations)					
ETD (low resolution) WCL phosphopeptides	82	56 (0)	4 (1)	5 (1)	17 (14)
ETD (high resolution) WCL phosphopeptides	54	40 (0)	3 (1)	1 (1)	10 (10)
ECD (high resolution) WCL phosphopeptides	41	29 (0)	3 (1)	1 (1)	7 (7)

[a] CID and ECD data were acquired with a 7T Thermo Finnigan LTQ-FT mass spectrometer. ETD data was acquired with a Thermo Fisher Orbitrap XL, using either the Orbitrap (high resolution) or LTQ (low resolution) detector. All the peptides included in the table contained multiple potential phosphorylation sites. The numbers in brackets indicate the number of errors or disagreements (or multiple top hits, when SLoMo was unable to localise the modification).

a hypothetical ion and a peak for a match may be customized, and is expressed in parts per million (ppm) rather than m/z. Expressing the fragment ion tolerance in ppm better reflects the errors in mass measurement in both low and high resolution data, i.e., Δppm is relatively constant across the m/z range of the mass spectrum, while $\Delta m/z$ increases with fragment m/z. Peptide scores are calculated using a binomial probability model. This approach however runs into computational problems as the number of trials increases and, above approximately 150 trials, it becomes impossible to calculate a score owing to overflow/underflow problems with the numbers used to calculate the binomial probability. (Overflow/underflow errors are caused by the computer trying to work with numbers which are too big/too small, respectively, for it to handle). When this situation occurs, SLoMo uses a Poisson model for calculations, with checks for number overflow or underflow fail, and an additional check being done to ensure that n × p and n × p × q are within 10% of each other, i.e., that the Poisson model is a close approximation to the equivalent binomial model. As for Ascore, the probabilities are converted into a log score via the equation:

$$\text{Score} = -10 \times \log(P)$$

As for Ascore, the peptide score for each site-localized peptide (i.e., peptide in which the modification is assigned to a specific site) is calculated as a weighted average of the scores for each peak depth (1 peak per window = 0.5; 2 = 0.75; 3 = 1; 4 = 1; 5 = 1; 6 = 1; 7 = 0.75; 8 = 0.5; 9 = 0.25; 10 = 0.25). Data from the two site-localized peptides with the highest peptide scores are used to generate the final score. At this stage only site-determining ions are used in the calculation. A site-determining ion is any ion unique to the site-localized peptide in question. The list of site-determining ions is compared against the list of peaks at x peaks per windows, where x is the lowest number peaks per window for which the difference between the two peptide scores is maximal. The score is calculated as described above and the final score is the difference between the score for the top and second placed modification sites. An example of the HTML output generated for each localization is shown in Figure 1.

Testing and Validation. SLoMo was tested against a series of data sets to ensure (i) it produced accurate site localizations for data which could also be analyzed by Ascore, and (ii) could generate accurate localization data for datasets generated by different dissociation techniques. Both synthetic phosphopeptides, with known sites of phosphorylation, and *in vivo* phosphorylated peptides with subsequent manual validation were used to assess SLoMo site localization.

Comparison with Ascore. In order to compare the output of SLoMo with that of Ascore, we took a set of 135 phosphopeptides available on the Ascore Web site, and analyzed them with both SLoMo and Ascore. From this set of 135, Ascore successfully (score ≥ 19) localized the site of phosphorylation in 79 peptides, compared to 75 using an identical score cutoff for SLoMo. The overlap between successfully localized phosphopeptides was approximately 90% (70 cases). For 8 of the 9 peptides which were confidently localized by Ascore but not by SLoMo, the SLoMo localization agreed but the SLoMo score was below 19. SLoMo could not distinguish between the top possibilities for the final peptide.

Within the set of peptides where either method produced a localization (regardless of score) (n = 116), a total of six were localized by SLoMo but not by Ascore, and 4 were localized by Ascore but not by SLoMo. There were also a total of three peptides where the site(s) of localization differed between the two algorithms, however the scores for these peptides were below 19 in all cases. This experiment also demonstrated the expected good correlation between Ascores and SLoMo scores (R^2 = 0.84, see Supplementary Figure 1, Supporting Information).

Application of SLoMo to ECD and ETD Data. To test the performance of SLoMo on ECD data, we acquired both CID and ECD mass spectra from two synthetic phosphopeptide libraries. Each library consists of a mixture of two thousand distinct phosphopeptides, with a common peptide backbone. Each library phosphopeptide has at least two potential sites of phosphorylation, with a maximum of five potential sites, depending upon the identity of the variable amino acids. The actual site of phosphorylation is fixed at serine 7 for both libraries. The sites assigned by SLoMo are shown in Table 1, top half. The performance of SLoMo was also tested on a small number of phosphopeptides enriched from mammalian whole-cell lysates, with manual validation of the assigned sites. Three data sets were used in these tests: ECD data and ETD data acquired with both high and low resolution. The results are shown in Table 1, bottom half. The error rates are within the expected ranges, that is <1.3%, 1.3−3.2% and 3.2−10% for scores of >19, 15−19, and 10−15, respectively.

Influence of Fragment Ion Mass Tolerance on Site Localization by SLoMo. The combined CID synthetic peptide libraries (Table 1, rows 1 and 2) were searched with varying

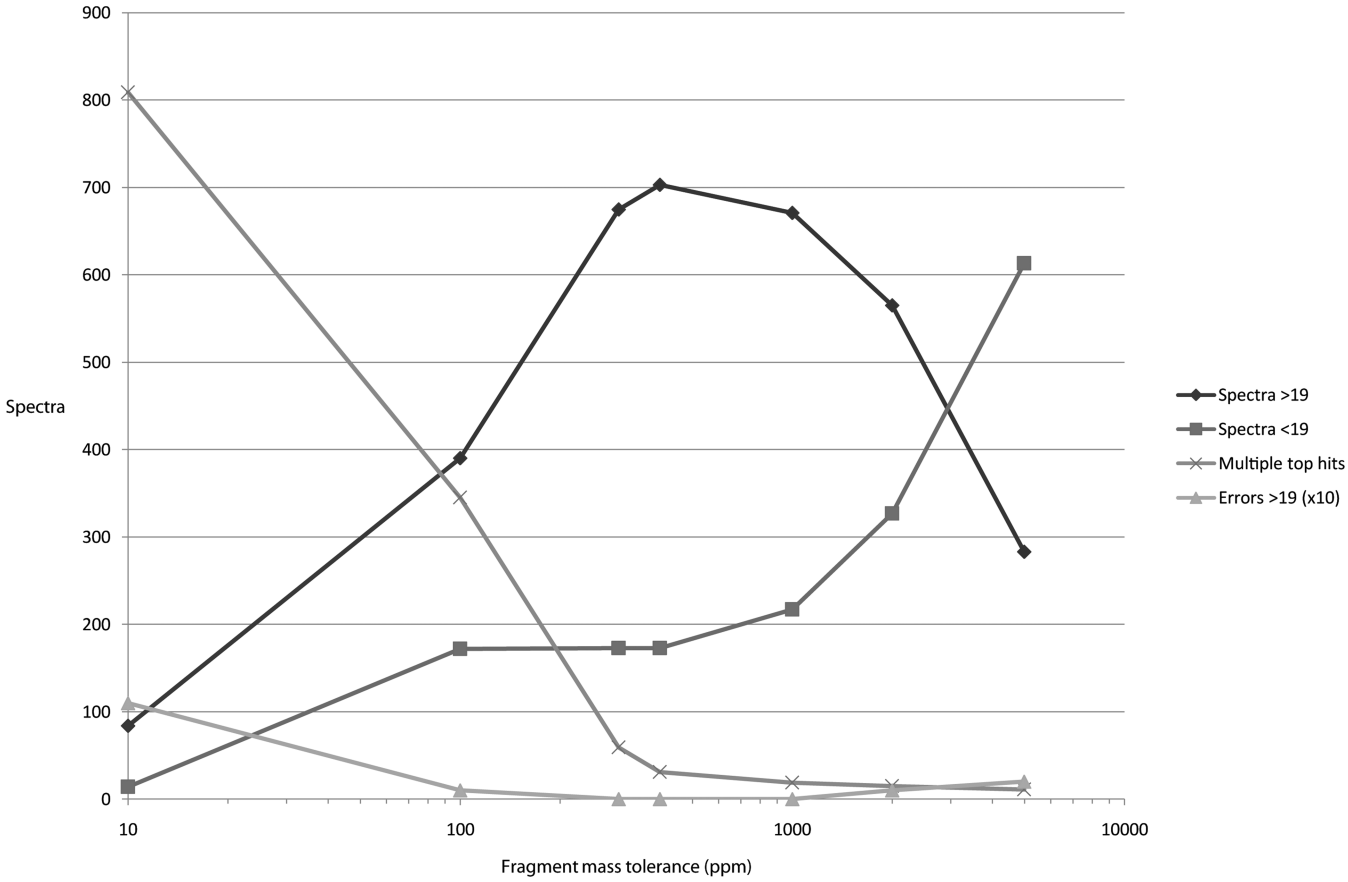


Figure 2. Influence of fragment mass tolerance on SLoMo localization of phosphorylation. The combined CID mass spectra ($n = 907$) of the synthetic phosphopeptide libraries were searched using different fragment mass tolerances. The number of errors >19 ranged from 11 to 0 and is shown multiplied 10-fold.

fragment ion tolerances, from 10 to 5000 ppm. The number of confident localizations reaches a maximum around the true mass accuracy of the acquired data (approximately 400 ppm error; **Figure 2**). Picking a mass tolerance much larger than necessary causes the SLoMo scores to drop, while reducing the mass tolerance to an unrealistically small value causes an increase in mass spectra with no site assignment (effectively a score of 0). The error rates were relatively constant, with the exception of a spike in high-scoring errors when the mass tolerance was reduced to 10 ppm.

Localization of Methionine Oxidation. To demonstrate the localization of a modification other than phosphorylation, a mouse whole-cell lysate sample was analyzed (the flow-through from the phosphopeptide enrichment described above) and sites of methionine oxidation were localized. In this case, the input to SLoMo was a pepXML file generated from a Sequest database search. An example of the successful localization of methionine oxidation is shown in Supplementary Figure 2 (Supporting Information).

Conclusion

The results presented here demonstrate that SLoMo is a suitable tool for localization of sites of modification within peptides identified by mass spectrometry. In this study, we demonstrate that phosphorylation can easily be localized using data obtained from a variety of fragmentation methods, namely ECD, ETD and CID. The high degree of concordance between SLoMo and Ascore and the low error rates returned from synthetic phosphopeptides and manual validation demonstrate the accuracy of the algorithm. SLoMo has been validated using both high and low resolution test data generated from two instruments (LTQ-FT and Orbitrap). We have also demonstrated the applicability of SLoMo to other modifications using the example of methionine oxidation, and shown that SLoMo accepts outputs from a variety of protein database search engines (OMSSA, Sequest).

SLoMo demonstrates general applicability to problems where localization of sites of modification is required for peptides identified in high throughput mass spectrometry experiments. With SLoMo, researchers now have a tool which is capable of undertaking site determination analysis for a range of different modifications examined using multiple dissociation techniques. User-defined ppm mass tolerances allow SLoMo to be applied to data generated from different instruments (e.g., ion-trap, QToF and FT-ICR) and the generic pepXML input format makes SLoMo compatible with multiple database search algorithms. The program is extensible and modifiable: end users can easily customize the parameters used to search for localizations. Similarly, adapting the program to accommodate any new dissociation technique which may become available is straightforward. SLoMo is available for download for multiple platforms from http://massspec.bham.ac.uk/slomo.

Acknowledgment. We acknowledge CRUK (C.M.B., D.L.C., J.K.H.), the EU FP6 Endotrack (S.M.M.S.), and the Wellcome Trust (074131) (H.J.C.) for funding.

Supporting Information Available: Supplementary Figure 1. Comparison of Ascore scores and SLoMo scores. Score output for 135 phosphopeptides examined using Ascore (*x*-axis) and SLoMo (*y*-axis).Supplementary Figure 2. SLoMo output showing successful localization of methionine oxidation. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mann, M.; Ong, S. E.; Gronborg, M.; Steen, H.; Jensen, O. N.; Pandey, A. *Trends Biotechnol.* **2002**, *20*, 261–8.
(2) Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Faergeman, N. J.; Mann, M.; Jensen, O. N. *Mol. Cell. Proteomics* **2005**, *4*, 310–27.
(3) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301–5.
(4) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2199–204.
(5) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2193–8.
(6) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. *Nat. Biotechnol.* **2006**, *24*, 1285–92.
(7) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127*, 635–48.
(8) Ruttenberg, B. E.; Pisitkun, T.; Knepper, M. A.; Hoffert, J. D. *J. Proteome Res.* **2008**, *7*, 3054–9.
(9) Wan, Y.; Cripps, D.; Thomas, S.; Campbell, P.; Ambulos, N.; Chen, T.; Yang, A. *J. Proteome Res.* **2008**, *7*, 2803–2811.
(10) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–33.
(11) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.
(12) Sweet, S. M. M.; Bailey, C. M.; Cunningham, D. L.; Heath, J. K.; Cooper, H. J. *Mol. Cell. Proteomics*, in press; DOI: 10.1074/mcp.M800451-MCP200.
(13) Sweet, S. M.; Cooper, H. J. *Expert Rev. Proteomics* **2007**, *4*, 149–59.
(14) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. *Mol. Syst. Biol.* **2005**, *1*, 2005–0017.
(15) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2004**, *4*, 1534–6.
(16) Cooper, H. J.; Hudgins, R. R.; Håkansson, K.; Marshall, A. G. *Int. J. Mass Spectrom.* **2003**, *228*, 723–728.
(17) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958–64.
(18) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. *Mol. Cell. Proteomics* **2005**, *4*, 873–86.
(19) Akbarzadeh, S.; Wheldon, L. M.; Sweet, S. M.; Talma, S.; Mardakheh, F. K.; Heath, J. K. *PLoS ONE* **2008**, *3*, e1873.
(20) Sweet, S. M.; Mardakheh, F. K.; Ryan, K. J.; Langton, A. J.; Heath, J. K.; Cooper, H. *J. Anal. Chem.* **2008**, *80*, 6650–7.
(21) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(22) Savitski, M. M.; Kjeldsen, F.; Nielsen, M. L.; Zubarev, R. A. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 113–20.

PR800917P